REMARKS

Claims 1-25 are all the claims presently pending in the application.

It is noted that the claim amendments are made only for more particularly pointing out the invention, and not for distinguishing the invention over the prior art, narrowing the claims or for any statutory requirements of patentability. Further, Applicant specifically states that no amendment to any claim herein should be construed as a disclaimer of any interest in or right to an equivalent of any element or feature of the amended claim.

Claims 1, 3, 5, 7, 9, 11, 13, 15, and 17-25 stand rejected under 35 U.S.C. § 103(a) as being unpatentable over U.S. Patent No. 5,895,470 to Pirolli, further in view of U.S. Patent Publication No. 2002/0165707 to Call.

Claims 2, 6, 10, 14, and 16 stand rejected under 35 U.S.C. § 103(a) as being unpatentable over Pirolli/Call, and further in view of U.S. Patent No. 5,950,189 to Cohen et al. Claims 4, 8, and 12 stand rejected under 35 U.S.C. § 103(a) as being unpatentable over Pirolli/Call/Cohen, and further in view of U.S. Patent No. 6,401,088 to Jagadish et al.

These rejections are respectfully traversed in the following discussion.

I. THE CLAIMED INVENTION

As described and claimed, for example, by claim 1, the present invention is directed to a method of converting a document corpus containing an ordered plurality of documents into a compact representation in memory of occurrence data.

A first vector is developed for the entire document corpus. The first vector comprises a listing of integers corresponding to terms in the documents such that each document in the document corpus is sequentially represented in the listing.

II. THE PRIOR ART REJECTIONS

The Examiner alleges that Pirolli, further in view of Call, renders obvious the present invention described by claims 1, 3, 5, 7, 9, 11, 13, 15, and 17-25 and, when further modified by Cohen, renders obvious the present invention defined by claims 2, 6, 10, 14, and 16. The Examiner further alleges that, when Pirolli, further in view of Call and Cohen, is then further

modified by Jagadish, the present invention described by claims 4, 8, and 12 is also rendered obvious.

Relative to the primary reference Pirolli, the rejection currently of record confused and/or contradictory. As best understood, the Examiner concedes that Pirolli has no teaching or suggestion of using integers or of developing a single listing of such integers for the entire document corpus.

Applicants submit that this concession alone, therefore, disqualifies Pirolli as serving as the primary reference for the prior art evaluation. That is, if Pirolli were to serve as the primary reference, the only portion of independent claims 1, 5, 9, 13, and 15 that would reasonably be satisfied by Pirolli would be the claim preamble.

Applicants submit that, as such, the principle of operation of Pirolli necessarily must be changed in order to meet the limitation of these independent claims, clearly a violation of MPEP§ 2143.01: "*If the proposed modification or combination of the prior art would change the principle of operation of the prior art invention being modified, then the teachings of the references are not sufficient to render the claims* prima facie *obvious.... The court reversed the rejection holding the "suggested combination of references would require a substantial reconstruction and redesign of the elements shown in [the primary reference] as well as a change in the basic principle under which the [primary reference] construction was designed to operate.*"

In the rejection for claims 5, 9, and 13 on page 3 of the Office Action, the Examiner points to lines 35-39 of column 7 of Pirolli, which read: "*An SCA engine processes text Wev pages by treating their contents as a sequence of tokens and gathering collection and document level object and token staistics (most notably token occurrence). A contiguous character string representing a word is an example of a token.*"

Applicants submit that this description, in combination with the description at lines 49-58 of column 7 have no suggestion of either using integers to represent the tokens or of treating the Web pages in any format except the conventional method based on each page being treated as a discreted document vector.

That is, there is no suggestion that the page-based document vectors be concatenated into a single vector. Applicants submit that such method to concatenate the vectors into a

single long vector representing all of the web pages on the web site would be significant and would require at least a suggestion in passing.

No such passing suggestion exists in Pirolli. That is, lines 53-60 of column 7 clearly describe a page-by-page approach: "*The token information is then used to create a document vector, where each component of the vector represents a word, step 403. Entries in the vector for a document indicate the presence or freqeuncy of a word in the document. The steps 401-403 are repeated for each Web page in the Web locality. For each pair of pages, the dot product of these vectors is computed, step 404.*"

Applicants submit that the above description fails to include any suggestion that the document vectors are to be concatenated into a single long vector representing the web site as a document corpus.

Nor does there seem to be a benefit in Pirolli for such single concatenated web site vector, since the purpose of the tokenization is that of calculating the similarity measure between pages, and it is this similarity measure that is carried forward for the remaining analysis using a matrix of vectors.

Moreover, Applicants submit that this matrix technique, even further described beginning line 4 of column 8 and demonstrated in Figure 5, clearly teaches against the very modification that the Examiner concedes to be missing in Pirolli of using a single concatenated vector that represents the tokenized page vectors of the pages in the web site.

Hence, turning to the clear language of the claims, in Pirolli, there is no teaching or suggestion of: "... developing a first vector for said entire document corpus, said first vector being a listing of integers corresponding to terms in said documents such that each said document in said document corpus is sequentially represented in said listing", as required by independent claim 1. The remaining independent claims have similar language.

Second, Applicants respectfully submit that, to one of ordinary skill in the art and contrary to the Examiner's assertion, Call does not overcome this basic deficiency in Pirolli. That is, Applicants submit that Call does not, in any way, describe that a single listing of integers is developed for an entire document corpus.

It is clear that Call is describing a technique that is intended to apply to each of separate documents, as demonstrated by the language in the final sentence of paragraphs [0024] and [0039], and the entirety of paragraphs [0106] and [0110].

That is, there is no suggestion whatsoever in Call to concatenate the integer listing of a first document with the integer listing of a second document. Applicants submit that, similar to the remark in the above discussion for Pirolli, if such concatenation feature were intended in Call, this feature would be too important to fail to mention explicitly, since the conventional wisdom, as clearly demonstrated in Figure 5 of Pirolli, is to treat documents as separate vectors. Applicants submit that Call fails to describe an alternate approach and, indeed, is not even oriented to a document corpus of documents.

As discussed in the final paragraph on page 13 of the disclosure, the advantage of these features of the present invention is particularly significant for a document corpus, such as the typical helpdesk document corpus discussed as a non-limiting example in the specification, having a million or more documents, the dictionary contains less than 32,000 terms, and each document contains less than a thousand words and has only one occurrence or a small number of occurrences of dictionary terms.

Neither Pirolli nor Call recognizes this specific type of document corpus.

Therefore, Applicants submit that, absent improper hindsight, there would be no motivation to modify either Pirolli or Call to use a single listing of integers for the entire document corpus.

Since this feature of a single uninterrupted listing of terms for the entire document corpus is a clear departure from the conventional wisdom, Applicants submit that the Examiner's initial burden has not been met until a very clear, explicit, and indisputable discussion in a prior art reference has been provided on the record.

The advantage of the technique of the present invention is that the occurrence data of the document corpus is totally preserved while being represented in "small sparse matrix vector form". The present inventors are unaware of any technique that uses less memory and still preserves the information content. The vector records in Pirolli do not provide a single vector representation of the entire database and do not provide a small sparse matrix vector format.

Relative to the rejection for claims 3, 7, and 11, Applicants submit that the Examiner's characterization that the alphabetical sorting described paragraph [0051] is not equivalent to the sorting of terms in each document so that identical terms are adjacent. Rather, this paragraph describes an alternate method used in the initial parsing of a document, described beginning in paragraph [0042] to determine whether a newly-identified token has occurred previously in the parsing.

Therefore, in this context of determining the first occurrence of terms and contrary to the Examiner's implication, the sorting in paragraph [0051] actually describes the alphabetical sorting of the terms in the T array. This is a different concept from that of sorting array T alphabetically. That is, the purpose of alphabetically sorting terms encountered in array T for purpose of easily comparing with a newly-identified token to determine whether the new token has already been used in array T would not require a replication of terms that are repeated in the array.

In contrast, in the present invention, the rearrangement of terms that occur more than once in a document is done for purpose of more easily normalizing the section of the concatenated vector that applies to that specific document.

Relative to the rejection for claims 2, 6, 10, 14, and 16, the Examiner relies upon Cohen. However, Applicants submit that, to one of ordinary skill in the art, the cited passage at lines 1-39 of column 11, merely describes in passing that normalized vectors are used. It does not, as the Examiner alleges, make any suggestion that the normalized vector is a third vector of a set of three vectors representing a document corpus. It is also pointed out that the "normalized vector" mentioned in these lines of Cohen is not the same concept as a vector "... comprising a sequential listing of floating point multipliers, each said floating point multiplier representing a document normalization factor", as required by the plain meaning of the claim language.

That is, Applicants submit that, to one of ordinary skill in the art, a sequential listing of normalization factors is not in any way equivalent to either a normalized vector or even a sequence of normalized vectors, or even, for that matter, the concept of a normalization factor.

Therefore, even if Cohen were to be somehow incorporated into Pirolli, it would not achieve the result described in these claims. Nor would it overcome the basic deficiency of Pirolli described above that a single concatenated vector be developed for an entire document corpus.

Jagadish is intended to address the specific formulation described in claims 4, 8, and 12, and, even assuming this reference to be properly combinable with Pirolli, does not overcome the deficiencies described above. Moreover, Applicants submit that the passage cited in Jagadish (e.g., lines 14-46 of column 8) relates to a probability of occurrence, a concept entirely different from the normalization factor of Cohen. Therefore, Applicants submit that modification of Cohen by Jagadish is not properly motivated by the description in the rejection currently of record.

Applicants also submit that the description in Jagadish fails to satisfy the plain meaning of the language in these claims, since neither Figure 6 nor the passage in column 8 provides the equation described in the claims.

Therefore, Applicants submit that, for all of the reasons above, the rejections currently of record fail to provide a *prima facie* rejection of the claimed invention. Applicants respectfully request that the Examiner reconsider and withdraw these rejections.

Therefore, claims 1-25 are clearly patentable over Pirolli, and the secondary references Call, Cohen, and Jagadish do not provide the elements necessary to complete the techniques of the present invention, let alone overcome the basic deficiencies of Pirolli.

As mentioned above, the present invention totally preserves the occurrence data of the document corpus while also providing the advantage of representing it in "small sparse matrix vector form". The prior art currently of record confirms the present inventors impression that no technique has yet been developed for a document corpus that uses less memory and still preserves the information content.

III. FORMAL MATTERS AND CONCLUSION

In view of the foregoing, Applicant submits that claims 1-25, all the claims presently pending in the application, are patentably distinct over the prior art of record and are in

condition for allowance. The Examiner is respectfully requested to pass the above application to issue at the earliest possible time.

Should the Examiner find the application to be other than in condition for allowance, the Examiner is requested to contact the undersigned at the local telephone number listed below to discuss any other changes deemed necessary in a telephonic or personal interview.

The Commissioner is hereby authorized to charge any deficiency in fees or to credit any overpayment in fees to Assignee's Deposit Account No. 09-0441.

Respectfully Submitted,

Date: 4/21/05

Frederick E. Cooperrider, Esq.
Registration No. 36,769

McGinn & Gibb, PLLC
8321 Old Courthouse Road, Suite 200
Vienna, VA 22182-3817
(703) 761-4100
Customer No. 48146